

Movie Box Office Prediction With Self-Supervised and Visually Grounded Pretraining

Qin Chao^{1,3}, Eunsoo Kim², and Boyang Li³

¹*Alibaba Group and the Alibaba-NTU Joint Research Institute, Singapore*

²*Nanyang Business School, Nanyang Technological University (NTU), Singapore*

³*School of Computer Science and Engineering, NTU, Singapore*

chao0009@e.ntu.edu.sg, {eunsoo, boyang.li}@ntu.edu.sg

Abstract—Investments in movie production are associated with a high level of risk as movie revenues have long-tailed and bimodal distributions [1]. Accurate prediction of box-office revenue may mitigate the uncertainty and encourage investment. However, learning effective representations for actors, directors, and user-generated content-related keywords remains a challenging open problem. In this work, we investigate the effects of self-supervised pretraining and propose visual grounding of content keywords in objects from movie posters as a pretraining objective. Experiments on a large dataset of 35,794 movies demonstrate significant benefits of self-supervised training and visual grounding. In particular, visual grounding pretraining substantially improves learning on movies with content keywords and achieves 14.5% relative performance gains compared to a finetuned BERT model with identical architecture.

Index Terms—Multimodal Learning, Self-supervised Learning, Visual Grounding, Box Office Prediction, Movie Revenue Prediction

I. INTRODUCTION

Movies are undoubtedly a preeminent form of art in the 21st-century human civilization. However, the business side of movie production is often less than glamorous. Statistics [1] show that box-office revenues have long-tailed and bimodal distributions, where a small number of movies take most of the profit and the majority barely make even. According to boxofficemojo.com¹, in 2019, the top-10 highest-grossing movies collected 13.2 billion US dollars or 37.4% of the global revenue of the top-200 movies. Three years into the pandemic, as of November 2022, the ratio balloons to 50.1%. The exorbitant risk of the industry drives producers to focus on superhero movies and sequels, whose outcomes are relatively predictable. Small studios that cannot afford to make high-budget movies share an ever smaller pie.

Algorithmic box office prediction holds the promise to help producers properly budget expenses, reduce risk, and encourage investment in creative and diverse content. The problem has attracted much research interest [2]–[15]. In this paper, we investigate the effects of self-supervised pretraining and visual grounding.

The star power of actors and directors is one of the most important factors determining box office revenue, but the data for each actor and director is limited. Even prolific directors typically make less than 30 movies throughout their

TABLE I: Examples of user-generated keywords from TMDB.

action, criminology, fbi, psycho, aircraft, robot
love, hate, high school, father-daughter relationship,
paris france, kingdom, based on novel or book

careers. Similarly, few modern actors play leading roles in more than 30 movies. By modern machine learning standards, these numbers are considered few-shot settings. To tackle data sparsity, we adopt self-supervised pretraining that encourages the network to learn the data distribution before training on box office data.

Another important, yet difficult to model, aspect of the box office is the movie content. The movie storyline is a complex artifact with multiple layers of semantics [16]–[19], which are challenging for even state-of-the-art AI to understand. To tackle this issue, we utilize user-generated content keywords from The Movie Database (TMDB)² to incorporate the movie content into the box office prediction problem. Table I shows example keywords. Compared to traditional genre categories, these keywords provide finer-grained categorization of content, including topic, plot, emotion, and even source-related information³.

To gain a precise understanding of these keywords, we further propose to ground the keywords in the visual modality — the movie posters. In the context of movies, the meaning of keywords can differ subtly from their daily usage. For example, the keyword *action* may be associated with explosion, car chasing, or martial art, deviating from its dictionary definition. The keyword *robot* typically refers to robots in science fiction or animation movies, rather than those on assembly lines. Recent research [20], [21] shows that grounding language in visual signals yields improved representation. In this paper, we find that this effect also exists and that the improved representation contributes to a better box office prediction. To our knowledge, this is the first paper that visually grounds textual information for box office prediction.

Overall, our research highlights the effectiveness of self-supervised pretraining and visual grounding in box office

²www.themoviedb.org

³More details are in the keyword contribution guide located at <https://www.themoviedb.org/bible/movie>

¹<https://www.boxofficemojo.com/year/world/2019/>

prediction. Our models relatively reduce prediction error by 7.8%~14.5% compared to the directly finetuned baseline BERT model under the same number of hyperparameters, whereas pretraining with visual grounding leads to up to 2.1% relative performance improvements.

With this paper, we make the following contributions. First, we propose self-supervised pretraining for movie box office forecasting that can utilize a combination of textual and numerical information. Second, we demonstrate that visual grounding user-generated keywords in movie posters significantly improves pretraining, suggesting a good correlation between movie content and the posters. Finally, we construct a large well-organized dataset for movie box office prediction and share it with the research community⁴.

II. RELATED WORK

Predicting Movie Success. Prior work has attempted to predict a number of indicators of commercial and artistic success, including the box office [2]–[4], return on investment [5]–[7], IMDb ratings [8], [9], critic reviews [11], and awards or award nominations [10]. Recently, with the advancement of ML, deep networks have begun to gain research attention [12]–[15].

In terms of features, aside from commonly adopted numeral features, [2], [5]–[7], [11] utilize textual features such as sentiment and topics. In particular, topics from Latent Dirichlet Allocation [5], [11] may be seen as a type of content feature. [9], [15] utilize fastText [22] and ELMO [23] word embeddings respectively. To our knowledge, the only prior work using visual features for box office prediction is [14], which incorporates movie poster features from a convolutional neural network during training. In contrast, our work leverages objects inside the poster to visually ground content keywords during pretraining, but do not use the poster during the finetuning stage.

Self-supervised Multimodal Pretraining. The success of pre-trained textual models such as BERT [24] has inspired a series of pretrained multimodal models [25]–[28], often adopting the masked language modeling (MLM) objective. Similar to a denoising autoencoder, the MLM objective trains the model to predict masked portions of the input. This seemingly simple training technique has demonstrated effectiveness across a wide range of downstream applications. Another line of work, such as CLIP [29] and BLIP [30], adopt a pretraining objective that distinguishes between correct image-text pairings and incorrect pairings.

A classic problem of cognitive science, the symbol grounding problem [31] is concerned with how words can gain their meaning as pointers to other concepts and objects. Computationally grounding textual tokens in visual images has demonstrated success in some applications [20], [21], [32]–[36]. In this work, we use movie posters as a source of visual grounding for the textual tokens — keywords. A movie poster is a widely used visual medium to advertise a movie long

before its release. Thus, we ground the tokens using objects from a single poster, and each token can be related to multiple objects and vice versa. Compared to the aforementioned prior work, which retrieve or generate relevant images for the textual descriptions, in our task correspondences between the keywords and the poster are not known *a priori* and must be discovered in a multi-instance manner.

III. METHODOLOGY

In this section, we first introduce the features used by the proposed network, followed by the pretraining strategies.

A. Features

We include both discrete features such as actors or directors and real-valued features such as movie budget. The embeddings of discrete tokens are learned from data. For real-valued features, we adopt prototype-based numeral embeddings [37]. Formally, the embedding function is formulated as $\text{NE}(x) : \mathbb{R} \rightarrow \mathbb{R}^D$ that maps a real number x to a D -dimensional vector with the component

$$\text{NE}_i(x) = \exp\left(-\frac{\|x - q_i\|_2}{\sigma^2}\right), \quad (1)$$

where $\{q_i\}_{i=0}^{D-1}$ are D evenly spaced numbers over a specified interval, e.g., $[-10, 10]$. Before applying the numeral embedding function, we normalize the values using logarithm or min-max normalization, depending on whether or not the feature has a long-tail distribution.

We broadly categorize the features used in forecasting box office revenue into four categories: investment & marketing, star power, content, and competition & seasonality.

Investment & Marketing. The production budget is often an indicator of the movie’s quality. Here we take the logarithm with base 10. Furthermore, we include the distributor company as a token as distributors with greater market power may release movies on more screens, which increases revenue.

Star Power. We include up to two directors, two writers, and three leading actors in our model. Each person is a unique token whose embeddings are trained from scratch. We also calculate the profitability of each person, which is defined as the average of the revenues of all previous movies that this person has participated in as one of the leading roles. Moreover, we incorporate the gender and age of each actor at the time of the movie release.

Movie Content. We first include genres and MPAA ratings. In addition, we also include an indicator for whether a movie is part of a franchise.

Inspired by the success of user-generated keywords as content descriptors [38], we collect user-generated keywords from TMDB, yielding a total of 7,700 unique keywords for 35,794 movies. Among the keywords, we observe many rare keywords and near-synonyms, which may hinder learning. For rare keywords, the lack of data may prevent accurate embedding estimation. Synonyms and near-synonyms cause problems for contrastive learning, which would force the

⁴<https://github.com/jdsannchao/MOVIE-BOX-OFFICE-PREDICTION>

An example of input with textual and numerical features:

[CLS][PG-13]1.5678[Genres][Action][Sci-Fi][Keywords][shield][superhero][Directors][Joss Whedon][Actors][Chris Evans][SEP]

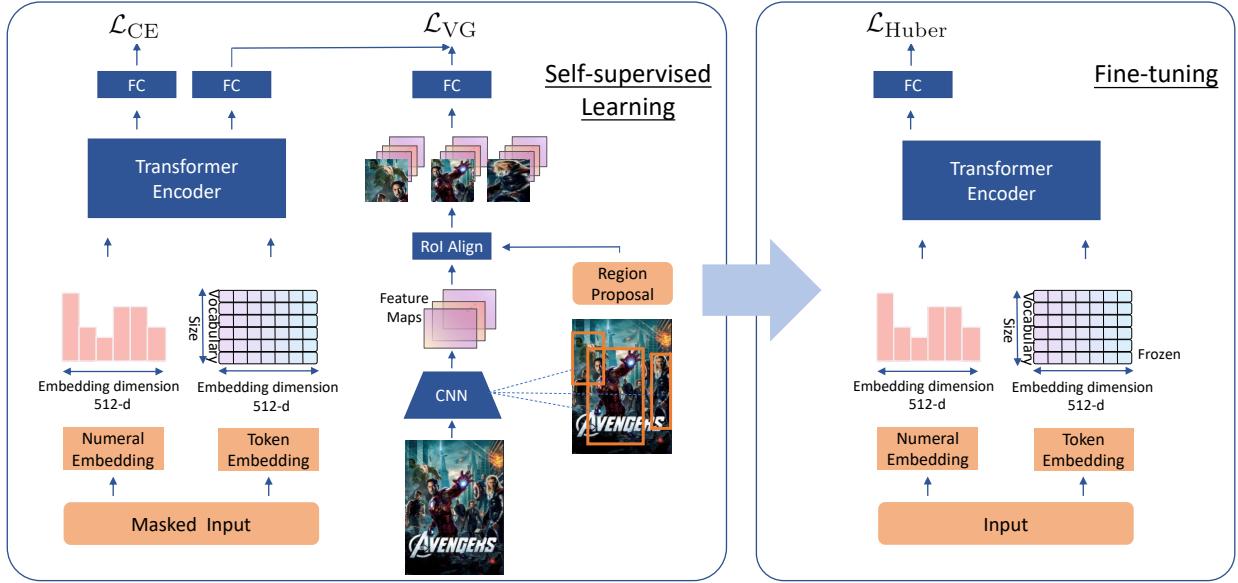


Fig. 1: The overall pipeline of self-supervised pretraining and finetuning on the box-office prediction task. The token embeddings are frozen during finetuning.

model to learn dissimilar embeddings for two words with similar meanings.

To overcome these issues, we cluster the keywords using both lexical similarity and co-occurrence statistics. To capture lexical information, we use 300-dimensional embeddings computed by fastText [22]. Next, we construct a movie-keyword term-frequency inverse-document-frequency (TF-IDF) matrix, which captures the co-occurrence statistics of keywords. From the TF-IDF matrix, we use the technique of [39] to construct embeddings for keywords. We extract the first 50 dimensions of the singular vectors to represent a keyword. The final representation is the 350-dimensional concatenation of the two vectors. We then perform average-link agglomerative clustering and use the resultant keyword clusters as features of movies. We show detailed cluster results in Appendix B.

Competition & Seasonality. To capture the effects of changing consumer tastes and holiday seasons, we include the year and month of the movie release as discrete tokens. Further, we model the competition intensity during the release window. We first identify competitors as those released two weeks before and after the current movie and have the same genre. After that, we sum up the overlap of content keywords, computed using the Jaccard similarity between the current movie and every competitor.

B. Self-supervised Pretraining

Figure 1 shows the overall pipeline. In the first stage, we pretrain a Transformer network on the MLM and visual grounding objectives. Next, we freeze the token embeddings and finetune the network on box-office prediction. We now introduce the pretraining tasks.

Masked Field Prediction. We adopt a pretraining objective similar to the masked language modeling task, which has been shown to be an effective pretraining method for natural language understanding [24] and multimodal understanding [25]. We randomly mask one token from each group of input features: genres, keywords, director/writer names, and actor names. The network is trained to predict the missing token. The prediction is formulated as cross-entropy losses, which we denote as \mathcal{L}_{CE} . By training the network to predict missing fields, we encourage the network to learn the correlations between the inputs, which could mitigate data scarcity issues.

Structured Visual Grounding. The content of the movie is undoubtedly crucial for box office, but understanding the user-generated content keywords is challenging. In particular, the content keywords may change in the context of motion pictures as the meaning of keywords can differ subtly from their daily usage as mentioned before.

We propose to ground the keywords in the visual modality provided by the movie posters. We conduct contrastive learning that encourages high similarity between a poster and the corresponding content keywords and suppresses the similarity between incorrectly paired posters and keyword sets. We first perform object detection on the poster with an off-the-shelf network, VinVL [40], but our method is not tied to this particular choice. We denote the extracted object features from the i^{th} movie as $\mathcal{Z}_i = \{z_m\}_{m=1}^M$. Note that we use the subscript i to denote the movie index. We also take the contextualized embeddings of the keywords from the output of the Transformer network, denoted as $\mathcal{X}_i = \{x_k\}_{k=1}^K$.

We define the similarity between the poster and the key-



Fig. 2: Multiple objects and keywords alignments for the movie *The Upside* (2019)

words as

$$\text{sim}(\mathcal{X}_i, \mathcal{Z}_i) = \sum_{(\mathbf{x}, \mathbf{z}) \in \mathcal{X}_i \times \mathcal{Z}_i} \exp\left(\frac{\mathbf{x}^\top \mathbf{z}}{\|\mathbf{x}\|_2 \|\mathbf{z}\|_2}\right), \quad (2)$$

where \times denotes the Cartesian product and $\|\cdot\|_2$ denote the L2 norm. To motivate the definition, we show one example poster and the associated keywords in Figure 2. We use colors of the keyword boxes to indicate cluster membership (e.g., “quadriplegia” and “handicapped” both belong to the red cluster). We observe that a cluster can correspond to multiple objects and one object may ground multiple clusters. For instance, the cluster “quadriplegia” is grounded by the wheelchair, the tire and the sitting man; the sitting man relates to the red and the purple clusters. Due to many-to-many relations, we follow [41] to define the similarity between the two sets as the sum of similarities of all possible pairs.

With randomly sampled negative pairs (i', j') , we define the visual grounding loss, \mathcal{L}_{VG} , as

$$\mathcal{L}_{VG} = -\frac{1}{N} \sum_{i=1}^N \log \left(\frac{\text{sim}(\mathcal{X}_i, \mathcal{Z}_i)}{\text{sim}(\mathcal{X}_i, \mathcal{Z}_i) + \sum_{(i', j')} \text{sim}(\mathcal{X}_{i'}, \mathcal{Z}_{j'})} \right) \quad (3)$$

where N is the total number of movies in the training set.

C. Finetuning on Box Office Prediction

In the finetuning stage, we train the network to predict box office revenues. We generate the prediction by feeding the average output from all input positions to a fully connected layer. Revenues follow a long-tailed distribution, which we approximate using a log-normal distribution. Hence, we take the base-10 logarithm of the revenue as the target value. To further reduce the effects of outliers, we train the network using the smooth L1 loss, also called the Huber loss,

$$\mathcal{L}_{\text{Huber}} = \begin{cases} 0.5(y - \hat{y})^2, & \text{if } |y - \hat{y}| < 1 \\ |y - \hat{y}| - 0.5, & \text{otherwise} \end{cases}, \quad (4)$$

where y is the ground truth and \hat{y} is the prediction.

TABLE II: Performance comparisons on the held-out box office test dataset. Our best model shows a 14.5% of accuracy improvement compared to BERT_{small}.

Model	Test Huber Loss(% improvement)	
Numerical features only		
Random Forest	0.3677	(-3.5%)
Textual and numerical features		
BERT _{small} finetuned	0.3553 _(baseline)	
BERT _{medium} finetuned	0.3446 _(2.5%)	
Our models		
	Clustering	Keywords
Random init.	0.3290 _(7.4%)	0.3265 _(8.1%)
+ MLM pretraining	0.3109 _(12.5%)	0.3133 _(11.8%)
+ VG pretraining	0.3070 _(13.6%)	0.3109 _(12.5%)
BERT embeddings init.	0.3137 _(11.7%)	0.3249 _(8.6%)
+ MLM pretraining	0.3102 _(12.7%)	0.3226 _(9.2%)
+ VG pretraining	0.3037 _(14.5%)	0.3182 _(10.4%)

IV. EXPERIMENTAL RESULTS

A. Data and Experimental Setup

We collect metadata of 35,794 movies from TMDB, including the period from 1920 to 2020. Total box office data for each movie during its original release period is crawled from IMDbPro. We use stratified sampling to divide the data into train, validation, and test sets in the ratios of 70/10/20, using “franchise movie” as the label for stratification. Using the method in §III-A, we cluster 7,700 keywords into 1,414 clusters. The number of clusters is tuned on the validation set. We use a 4-layer Transformer, with model dimension $d_{\text{model}} = 512$, fully connected layer dimension $d_{ff} = 512$, and 4 attention heads. The architecture is the same as BERT_{small}. More hyperparameters are reported in the Appendix A.

B. Baselines

We introduce three types of baseline models. The first is a Random Forest (RF). We feed only numerical features to the RF as one-hot encodings of the discrete features would have too many dimensions. Next, we introduce pretrained BERT models of small and medium sizes and finetune them on box office prediction. To mimic the classic BERT input, we concatenate all the input tokens into one sentence, while rounding numeral features to one decimal point, and then apply the BERT tokenizer. Lastly, we compare against a random initialized BERT_{small} directly trained on box office prediction and a BERT_{small} with pretrained BERT embeddings for actors, crew members, genres and keywords. When a name contains multiple words, we use the average of the pretrained BERT embeddings. For a keyword cluster, we use the embedding of the keyword in the cluster appearing the most frequently.

C. Results and Discussion

In Table II, we report the test-set Huber loss for all models, as well as their performance relative to BERT_{small}. Pretrained BERT models easily outperform the RF baseline,

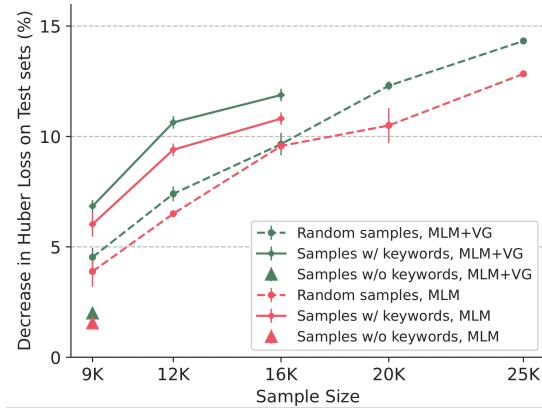


Fig. 3: Test losses on box office prediction with different training set sizes. Vertical bars indicate standard deviations. Exact numbers are reported in Appendix E.

but are inferior to the MLM and VG pretraining. Although the larger BERT_{medium} outperforms BERT_{small}, it underperforms our MLM-pretrained networks by more than 10% relatively. The domain gap between movie and textual data used in pretraining and our feature engineering likely contribute to the performance gaps.

Notably, VG pretraining obtains sizeable improvements on top of MLM for both types of embedding initialization. The fact that VG pretraining leads to improvement even with BERT-pretrained token embeddings corroborates our hypothesis that keywords may have specialized meanings in the movie context and visual grounding may help capture the specialized semantics. Finally, the best test loss of 0.3037, or 14.5% improvement relative to BERT_{small}, is achieved by MLM+VG pretraining.

Content Keywords and Scaling. As not all movies come with user-supplied keywords, we further investigate the effects of pretraining on movies with and without content keywords. We split the training set into movies with keywords (16K out of 25K) and movies without (9K out of 25K). As comparison baselines, we also create random subsets of the entire training set of sizes 9K, 12K, 16K, 20K, and 25K. We report losses on the same test set when the MLM+VG network is training on different training sets in Fig. 3. We note that with equal amount of training data, MLM and VG both exhibit stronger generalization when training on data with keywords than randomly mixed data. This agrees with our intuition as MLM exploits correlation between keywords and VG further reinforces keywords with visual information. In Fig. 6 in the Appendix, we examine if VG improves upon MLM for movies with keywords. We observe that the improvement of MLM+VG over MLM widens as training data increase, suggesting VG scales well and its effectiveness grows with data.

Effects of Keywords Clustering. We examine the effects of keyword clusters. Table II compares results with keyword clustering ("Clustering") with those on raw keywords ("Key-

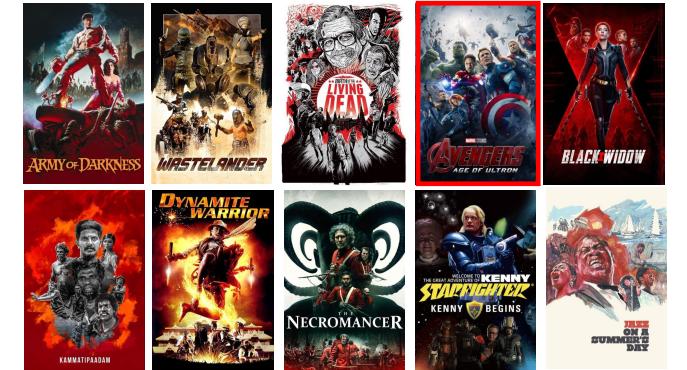


Fig. 4: **Top:** Retrieved posters from the keyword “love” in the context of a romantic movie, *One Day* (2009); **Bottom:** Retrieved posters using the keyword “superhero” in the context of *The Avengers* (2012).

words”). In most cases, keyword clusters provide performance gains, especially when pretrained BERT embeddings are used. A possible reason is that near-synonyms have similar BERT embeddings that are difficult for the model to differentiate, and clusters alleviate this problem.

D. Poster Retrieval Examples

We qualitatively examine the effects of visual grounding. Figure 4 shows posters that are most similar to keywords within the contexts of movies. The top two rows are retrieved for the keyword “love” in the context of a romantic movie *One Day* (2009). The majority of posters fall under the romance genre and visualize a couple embracing one another. The bottom ten posters are retrieved for the keyword “superhero” in the movie *The Avengers* (2012). The results are mostly action movies with a hero at the center of the poster surrounded by others. Appendix E contains more examples.

V. CONCLUSION

Box office revenue is influenced by a plethora of entangled factors that are often hard to observe, let alone computationally model. An important challenge in box office prediction is hence to learn representations that capture movie semantics and correlate well with the target variable. For this purpose,

we propose to pretrain a transformer network with masked language modeling and visual grounding objectives, which demonstrate substantial performance boost. We hope these results could inspire subsequent research on multimodal box-office prediction.

VI. ACKNOWLEDGMENTS

This research is supported, in part, by Alibaba Group through the Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (No. ANGC-2020-011), the National Research Foundation Fellowship (No. NRF-NRFF13-2021-0006), Singapore, and NTU Start-Up Grant.

REFERENCES

- [1] Pan, R. & Sinha, S. The statistical laws of popularity: universal properties of the box-office dynamics of motion pictures.. *New Journal Of Physics*. (2010)
- [2] Apala, K., Jose, M., Motnam, S., Chan, C., Liszka, K. & Gregorio, F. Prediction of movies box office performance using social media. *2013 IEEE/ACM International Conference On Advances In Social Networks Analysis And Mining (ASONAM 2013)*. pp. 1209-1214 (2013)
- [3] Parimi, R. & Caragea, D. Pre-release box-office success prediction for motion pictures. *International Workshop On Machine Learning And Data Mining In Pattern Recognition*. pp. 571-585 (2013)
- [4] Simonoff, J. & Sparrow, I. Predicting movie grosses: Winners and losers, blockbusters and sleepers. *Chance*. **13**, 15-24 (2000)
- [5] Lash, M. & Zhao, K. Early predictions of movie success: The who, what, and when of profitability. *Journal Of Management Information Systems*. **33**, 874-903 (2016)
- [6] Eliashberg, J., Hui, S. & Zhang, Z. Assessing box office performance using movie scripts: A kernel-based approach. *IEEE Transactions On Knowledge And Data Engineering*. **26**, 2639-2648 (2014)
- [7] Rajput, P., Sapkal, P. & Sinha, S. Box office revenue prediction using dual sentiment analysis. *International Journal of Machine Learning And Computing*. **7**, 72-75 (2017)
- [8] Cizmeci, B. & Öğüdücü, Ş. Predicting IMDb ratings of pre-release movies with factorization machines using social media. *2018 3rd International Conference On Computer Science And Engineering (UBMK)*. pp. 173-178 (2018)
- [9] Shafaei, M., Lopez-Monroy, A. & Solorio, T. Exploiting Textual, Visual, and Product Features for Predicting the Likeability of Movies. *FLAIRS*. (2019)
- [10] Boccardelli, P., Brunetta, F., Vicentini, F. & Others. What is critical to success in the movie industry? A study on key success factors in the Italian motion picture industry. *DIME*. (2008)
- [11] Kim, E., Ding, M., Wang, X. & Lu, S. Does Topic Consistency Matter? A Study of Critic and User Reviews in the Movie Industry. *Journal Of Marketing*. (2023)
- [12] Quader, N., Gani, M. & Chaki, D. Performance evaluation of seven machine learning classification techniques for movie box office success prediction. *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*. pp. 1-6 (2017)
- [13] Antipov, E. & Pokryshevskaya, E. Are box office revenues equally unpredictable for all movies? Evidence from a random forest-based model. *Journal Of Revenue And Pricing Management*. **16**, 295-307 (2017)
- [14] Zhou, Y. & Yen, G. Evolving Deep Neural Networks for Movie Box-Office Revenues Prediction. *2018 IEEE CEC*. (2018)
- [15] Kim, Y., Cheong, Y. & Lee, J. Prediction of a movie's success from plot summaries using deep learning models. *Proceedings of The Second Workshop on Storytelling*. pp. 127-135 (2019)
- [16] Barot, C., Potts, C. & Young, R. A tripartite plan-based model of narrative for narrative discourse generation. *AIIDE*. (2015)
- [17] Cutting, J. Narrative theory and the dynamics of popular movies. *Psychonomic Bulletin & Review*. **23**, 1713-1743 (2016)
- [18] Li, B. Learning knowledge to support domain-independent narrative intelligence. Ph.D. Dissertation. Georgia Institute of Technology. 2015.
- [19] Davis, N., Li, B., O'Neill, B., Riedl, M. & Nitsche, M. Distributed creative cognition in digital filmmaking. *Proceedings of The 8th ACM Conference on Creativity and Cognition*. (2011)
- [20] Tan, H. & Bansal, M. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. *EMNLP*. (2020)
- [21] Lu, Y., Zhu, W., Wang, X., Eckstein, M. & Wang, W. Imagination-Augmented Natural Language Understanding. *NAACL*. (2022)
- [22] Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching Word Vectors with Subword Information. *ArXiv Preprint arXiv:1607.04606*. (2016)
- [23] Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefanik, I., Jarkiewicz, M. & Okruszek, L. Detecting formal thought disorder by deep contextualized word representations. *ArXiv Preprint ArXiv: 1802.05365*. (2018)
- [24] Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint arXiv:1810.04805*. (2018)
- [25] Lu, J., Batra, D., Parikh, D. & Lee, S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*. **32** (2019)
- [26] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F. & Dai, J. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv:1908.08530*. (2019)
- [27] Tan, H. & Bansal, M. LXBERT: Learning Cross-Modality Encoder Representations from Transformers. *EMNLP-IJCNLP*. (2019)
- [28] Huang, Z., Zeng, Z., Liu, B., Fu, D. & Fu, J. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *ArXiv Preprint arXiv:2004.00849*. (2020)
- [29] Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. & Others Learning transferable visual models from natural language supervision. *ICML*. pp. 8748-8763 (2021)
- [30] Li, J., Li, D., Xiong, C. & Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *ArXiv Preprint 2201.12086*. (2022)
- [31] Harnad, S. The symbol grounding problem. *Physica D: Nonlinear Phenomena*. **42**, 335-346 (1990).
- [32] Yang, Y., Yao, W., Zhang, H., Wang, X., Yu, D. & Chen, J. Z-LaVi: Zero-Shot Language Solver Fueled by Visual Imagination. *ArXiv Preprint 2210.12261*. (2022)
- [33] Zhang, L., Chen, Q., Siebert, J. & Tang, B. Semi-Supervised Visual Feature Integration for Language Models through Sentence Visualization. *ICMI*. (2021)
- [34] Kiros, J., Chan, W. & Hinton, G. Illustrative Language Understanding: Large-Scale Visual Grounding with Image Search. *ACL*. (2018)
- [35] Liu, X., Yin, D., Feng, Y. & Zhao, D. Things not Written in Text: Exploring Spatial Commonsense from Visual Signals. *ACL*. (2022)
- [36] Long, Q., Wang, M. & Li, L. Generative imagination elevates machine translation. *ArXiv Preprint arXiv:2009.09654*. (2020)
- [37] Jin, Z., Jiang, X., Wang, X., Liu, Q., Wang, Y., Ren, X. & Qu, H. Numgpt: Improving numeracy ability of generative pre-trained models. *ArXiv Preprint 2109.03137*. (2021)
- [38] Annalyn, N., Bos, M., Sigal, L. & Li, B. Predicting personality from book preferences with user-generated content labels. *IEEE Transactions on Affective Computing*. **11**, 482-492 (2018)
- [39] Zhang, Y., Li, B., Liu, Y., Wang, H. & Miao, C. Initialization matters: regularizing manifold-informed initialization for neural recommendation systems. *ACM SIGKDD*. (2021)
- [40] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y. & Gao, J. Vinvl: Revisiting visual representations in vision-language models. *CVPR*. (2021)
- [41] Miech, A., Alayrac, J., Smaira, L., Laptev, I., Sivic, J. & Zisserman, A. End-to-end learning of visual representations from uncurated instructional videos. *CVPR*. pp. 9879-9889 (2020)
- [42] He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. *JCCV*. (2017)

APPENDIX

A. Experimental Setup and Hyperparameters

During the visual grounding pretraining, we randomly select up to 6 keywords per movie and up to 20 objects per poster to compute the similarity. The feature maps of each object have a dimension of $2048 \times 4 \times 4$, it is the output from VinVL after the ROI Align [42] and an adaptive average pooling layer. After that, the feature map is then flattened spatially and linearly projected to $\mathbb{R}^{d_{model}}$, where $d_{model} = 512$.

We use a batch size of 2048 when pretraining the model under MLM objective and reduce the size to 326 when adding the VG objective. The learning rate is $3e-4$. The optimizer we used is Adam with weight decay equals to $1e-4$. During the fine tune stage, we search for the best performance on the validation dataset in the combinations of learning rate in $[1e-3, 3e-4, 1e-4]$ and batch size in $[328, 512, 1024]$.

B. Clustering Samples and the Number of Unique Tokens

TABLE III: Some examples of the clustering results. The representative word of a cluster is the most frequent keyword in this set.

Cluster	Elements
love	'love', 'loved', 'hate', 'unhappy', 'waiting', 'happy', 'grateful', 'lucky', 'expecting', 'loving'
superhero	'superhero', 'villainess', 'villain', 'symbiote', 'sidekick', 'superhuman', 'teamup', 'nemesis', 'superheroes', 'supervillain'
psycho	'psycho', 'psychotic', 'pyromaniac', 'psychopathic', 'homicidal', 'deranged'

TABLE IV: # unique tokens for each textual features

Feature Name	Example	No. Unique Tokens
Release Year	release year range from 1920-2020	100
Release Month	12 tokens for 12 months	12
MPAA	G, PG, PG-13, R, NC17, NotRated, N.A.	7
Production Company	group small studios (produced less than 10 movie) into 'Others'	594
Distributor Company	group small studios (produced less than 10 movie) into 'Others'	407
Franchise	yes or no	2
Copycat	yes or no	2
Genres	e.g. Drama, Romance	18
Keywords	e.g. 'friendship'	1414
Crew Names	e.g. 'Steven Spielberg' is a token	15333
Cast Names	e.g. 'Leonardo DiCaprio' is a token	20366

C. Meta Data From TMDB

```

{'adult': False,
'backdrop_path': 'c1BaOxC8bo5ACFYkYYxL0bBWRaq.jpg',
'belongs_to_collection': None,
'budget': 4000000,
'genres': [{'id': 80, 'name': 'Crime'}, {'id': 35, 'name': 'Comedy'}],
'homepage': 'https://www.miramax.com/movie/four-rooms/',
'id': 5,
'imdb_id': 'tt0113101',
'original_language': 'en',
'original_title': 'Four Rooms',
'overview': "It's Ted the Bellhop's first night on the job ...and the hotel's very unusual guests are about to place him in some outrageous predicaments. It seems that this evening's room service is serving up one unbelievable happening after another.",
'popularity': 15.811,
'poster_path': '75aHniNOYXh4M7L5shoeQ6NGykP.jpg',
'production_companies': [{"id": 14,
'logo_path': 'm6AHu84oZQxvq7n1rsvMNJIAsMu.png',
'name': 'Miramax',
'origin_country': 'US'},
{'id': 59,
'logo_path': 'yH70MeSxhfP0AVM6iT0rsF3F4ZC.png',
'name': 'A Band Apart',
'origin_country': 'US'},
{'production_countries': [{"iso_3166_1": "US",
'name': "United States of America"}],
'release_date': '1995-12-09',
'revenue': 4257354,
'runtime': 98,
'spoken_languages': [{"english_name": "English",
'iso_639_1': "en",
'name': 'English'}],
'status': 'Released',
>tagline': "Twelve outrageous guests. Four scandalous requests. And one lone bellhop, in his first day on the job, who's in for the wildest New year's Eve of his life."}],
'title': 'Four Rooms',
'video': False,
'vote_average': 5.7,
'vote_count': 2146}

{'id': 5,
'keywords': [{"id": 612, 'name': 'hotel'},
{'id': 613, 'name': "new year's eve"},
{'id': 616, 'name': 'witch'},
{'id': 622, 'name': 'bet'},
{'id': 922, 'name': 'hotel room'},
{'id': 2700, 'name': 'sperm'},
{'id': 9706, 'name': 'anthology'},
{'id': 12670, 'name': 'los angeles, california'},
{'id': 160488, 'name': 'hoodlum'},
{'id': 187056, 'name': 'woman director'}]}

{'id': 5,
'cast': [{"adult': False,
'gender': 2,
'id': 3129,
'known_for_department': 'Acting',
'name': 'Tim Roth',
'original_name': 'Tim Roth',
'popularity': 15.779,
'profile_path': '/qSzF2i9gz6c6DbAC5RoIq8sVqX.jpg',
'cast_id': 42,
'character': 'Ted the Bellhop',
'credit_id': '52fe420dc3a36847f80001b7',
'order': 0},
...
'crew': [{"adult': False,
'gender': 1,
'id': 3110,
'known_for_department': 'Directing',
'name': 'Allison Anders',
'original_name': 'Allison Anders',
'popularity': 0.6,
'profile_path': '/ln8nIx6UjxpMLVQ1StCJpx6fyL7.jpg',
'credit_id': '52fe420dc3a36847f800012d',
'department': 'Directing',
'job': 'Director'}]}

```

```

{'adult': False,
'gender': 1,
'id': 3110,
'known_for_department': 'Directing',
'name': 'Allison Anders',
'original_name': 'Allison Anders',
'popularity': 0.6,
'profile_path': '/In8nIx6UjxpMLVQ1StCJpx6fyL7.jpg',
'credit_id': '52fe420dc3a36847f80001c9',
'department': 'Writing',
'job': 'Writer'}

```

D. Model Input

Feature Name	Tokens / Values
release_year ((is the input of both RF model and Transformer model))	2012
release_month (both)	March
MPAA (both)	PG-13
Budgets (both)	7.892094608
producer (both)	Lionsgate
distributor (both)	Lionsgate
N_competitors (both)	0.693147181
competitor_similarity (both)	0
franchise (both)	Yes
collection name	The Hunger Games_0
N_person (both)	0.693147181
N_man (both)	0
N_woman (both)	0
genres (both)	[genres] Adventure Fantasy Science Fiction
clusters	[clusters] retelling socialism backgammon interpretation [Directors] Gary Ross
Directors	1.098612289
Director1 experience (both)	0.875335786
Director1 profitability (both)	
same for Director2	
Writers	[Writers] Billy Ray
Writer1 experience (both)	1.386294361
Writer1 profitability (both)	0.86888262
similar for Writer2	Gary Ross 0 0
Actors	[Actors] Jennifer Lawrence
Actor1	Female
Actor1 Gender	
Actor1 Age	22
Actor1 experience (both)	0.693147181
Actor1 profitability (both)	0.792347251
same for Actor2, Actor3	Josh Hutcherson Male 20 1.098612289 0.928510042 Liam Hemsworth Male 22 0 0

E. More Experimental results

Adjust Loss Weights. Figure 5 shows the impact of the VG loss weight during pretraining on test Huber loss in the fine-tuning stage. That is, we vary the weight of VG loss and MLM

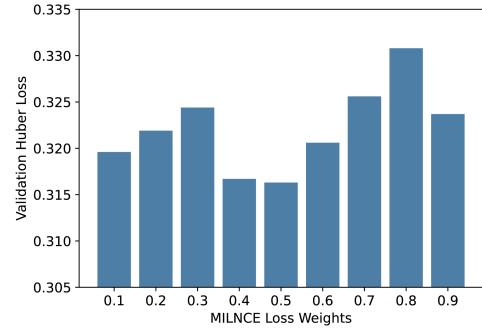


Fig. 5: Analysis on VG loss weights when pretrain the model on MLM and VG objective jointly.

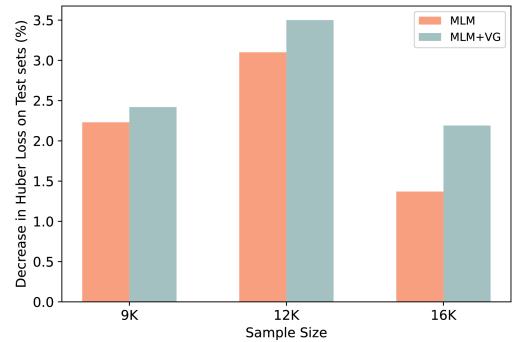


Fig. 6: Huber loss decrease between training on data with keywords and random data, for both MLM pretraining model and MLM+VG pertaining model with sample size vary. The gap between MLM and MLM+VG grows as the sample size increases.

loss to examine its impact on the test Huber loss. Based on our experiment, setting the equal weight to the MLM and VG loss attains the lowest test Huber loss. The results suggest that the movie context information and visual grounding equally contribute to the keyword representations, showcasing the non-negligible impact of visual grounding.

Detailed Comparison between MLM and MLM+VG. In Table V, we show detailed numbers comparing the two pre-training objectives, MLM and MLM+VG. For each model, we run three random trials and then compute the average and standard deviation. These numbers are reflected in Fig. 3 in the main text.

In Fig. 6, we show relative test loss improvement of training on data with keywords over training on randomly sampled data, which contain movies with keywords and without keywords. The green bars indicate the average improvements of the MLM+VG model and the red bars indicate the average improvements of the MLM model. We note that as the sample size increases, the gap between the MLM+VG model and the MLM-only models increases, suggesting that the effectiveness of visual grounding increases with data.

TABLE V: For each model, we run three random trials and then compute the average and standard deviation

sample size	Average Test Huber Loss(std.)	
	MLM pretraining	MLM+VG pretraining
random samples	25k	0.3097 _(0.0004)
	20k	0.3180 _(0.0028)
	16k	0.3213 _(0.0010)
	12k	0.3322 _(0.0002)
	9k	0.3415 _(0.0025)
with keywords	16k	0.3169 _(0.0012)
	12k	0.3219 _(0.0007)
	9k	0.3339 _(0.0020)
w/o keywords	9k	0.3498 _(0.0014)
		0.3482 _(0.0002)



Fig. 9: Poster retrieval using the keyword ‘friendship’ in the context of a typical Family & Animation movie *Toy Story* (1995)



Fig. 7: Poster retrieval using the keyword ‘psycho’ in the context of a typical Thriller movie *The Silence of the Limb* (1991)



Fig. 8: Poster retrieval using the keyword ‘war’ in the context of a typical War movie *Saving Private Ryan* (1998)